多示例多标签学习在中文专利自动分类中的应用研究*

■ 包翔[1] 刘桂锋[1] 崔靖华[2]
[1] 江苏大学科技信息研究所 镇江 212013 [2] 南京大学信息管理学院 南京 210093

摘 要：[目的/意义]旨在对大量的中文专利实现快速分类,满足专利审查以及情报分析等工作的要求。[方法/过程]结合专利文本的固有格式以及存在多个IPC分类号的实际情况,将多示例多标签学习应用于专利自动分类中,在介绍几种经典的多示例多标签模型的基本原理之后,将这些模型运用于中文专利IPC分类号的确定。[结果/结论]实验证明,多示例多标签模型适合运用在专利的自动分类中,并且从Average precision、Hamming Loss、Ranking Loss、One Error、Coverage、Training time等指标分析可以发现,MIMLRBF模型能快速、准确地运用在中文专利IPC分类号的确定中,为大规模专利的自动分类提供借鉴。

关键词：专利 分类 IPC分类号 多示例多标签
分类号：G251
DOI：10.13266/j.issn.0252-3116.2021.08.011

1 引言

专利作为极其重要的产权激励工具,为科技创新市场化保驾护航[1],近年来,我国的发明专利申请量均居世界首位,如此庞大的专利申请数量体现了我国科技实力的增强,同时也对专利的管理、分析、审查等方面提出了更高的要求。

专利分类是对海量专利文献组织、检索、分析和管理的有效手段,目前国际上应用广泛的专利分类体系包括国际专利分类IPC、美国专利分类USPC、欧洲专利分类ECLA、日本专利分类FI/F-term和联合专利分类CPC等[2],依据以上的体系进行专利分类使得专利的快速检索、定位成为可能。但是,现阶段的专利分类号的确定主要依靠人工判断,存在受标注人知识结构影响等弊端,因此,引入智能化技术解决专利的分类问题对于提升分类效率和准确率具有重要的意义。

专利文本自动化分类系统主要包含两方面的研究[3]:一是专利文本的特征提取算法;二是专利文本分类算法。在特征提取方面应用最为广泛的专利特征提取算法是词袋法(Bag of Words, BOW)和词频-反向词频(Term Frequency-Inverse Document Frequency, TF-IDF),但是两种模型都舍弃了文本中大量的信息,因此词向量(Word Embedding)[3]开始受到关注,较为经典的词向量模型有连续词袋模型(Continuous Bag of Words, CBOW)与Skip-Gram模型[4];温超东等[5]基于ALBERT预训练的动态词向量代替传统Word2vec等方式训练的静态词向量,提升了词向量的表征能力;余本功等[6]将专利文本分别映射为Word2vec词向量序列和POS词性序列,使用两种特征通道训练模型。在专利文本分类方法研究领域,传统的机器学习方法经常被用在专利分类中,包括朴素贝叶斯算法(Naive Bayesian, NB)、K最近邻(K-Nearest Neighbor, KNN)、支持向量机(Support Vector Machine, SVM)、逻辑回归(Logistics Regression, LR)、协同过滤技术等[7-8];近年来,深度学习技术也被广泛应用在专利分类中,其中有基于神经网络的方法,例如基于卷积神经网络(Convolutional Neural Network, CNN)[9]、双向门控循环单元(Bidirectional Gating Recurrent Unit, BiGRU)[5]、门控循环单元(Gated Recurrent Unit, GRU)[6]等神经网络技术,还有学者将其与其他方法进行结合,例如周成等[10]基于自组织映射(Self-Organizing Feature Map, SOM)和SVM的专利分类模型使用自组织映射方法确

* 本文系江苏省高校哲学社会科学研究一般项目"主题模型在高校图书馆知识产权信息服务中的研究与实践"(项目编号:2019SJA1870)和江苏省高校自然科学研究面上项目"基于多示例多标签学习及深度神经网络的专利主题分类研究"(项目编号:19KJB520005)研究成果之一。
作者简介：包翔(ORCID:0000-0002-2233-5739),馆员,硕士,E-mail:bx425bob@163.com;刘桂锋(ORCID:0000-0002-7209-3862),研究馆员,博士;崔靖华(ORCID:0000-0001-9723-3414),博士研究生。
收稿日期:2020-10-28 修回日期:2021-01-13 本文起止页码:107-113 本文责任编辑:王传清

定专利类别,采用随机森林(Random Forest, RF)进行重要性排序和特征选择;Y. Lu 等[11]利用长短期记忆网络(Long Short-Term Memory, LSTM)与基于注意力(Attention)的双向 LSTM 相结合形成模型训练专利语料,通过 Softmax 分类模型进行分类;吕璐成等[2]基于 Word2Vec、CNN、循环神经网络(Recurrent Neural Network, RNN)、Attention 机制等深度学习技术,在考虑专利文本语序特征、上下文特征以及分类关键特征的前提下,设计了 7 种深度学习模型。

上述方法从各个角度对专利分类进行了研究,也取得了不错的效果,但是这些方法未考虑到专利文本层次结构明显、主题描述规范、包含有多个分类号等特有的结构。近年来,多示例多标签(Multi-Instance Multi-Label learning, MIML)学习是一种发展极其迅速的机器学习模型,在文本、图像分类中取得不错的效果[12-15]。因此,本研究结合专利文本的固有格式以及存在多个分类号的实际情况,研究基于 MIML 的专利自动分类方法并进行评价。

2 MIML 介绍

2.1 MIML 概念以及专利文本结构

MIML 是一种新型的机器学习模型,它与其他机器学习方法不一样的地方在于训练集不是由若干示例组成,它是由一组含有标签的包(Bag)组成,若干示例构成了一个包,而且一个包可以对应一个或者多个标签,若一个包中至少存在一个该标签的正例,则这个包有对应的标签,若一个包中不存在该标签的正例,则该包没有对应的标签。如图 1 所示:

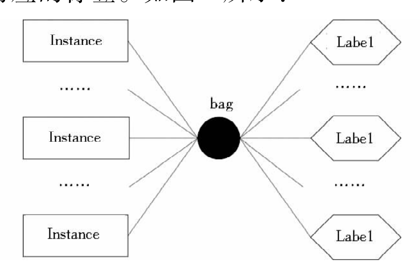


图 1　多示例多标签学习

MIML 通过对已经标定类别的包来建立学习模型,然后根据该模型预测未知包的所属标签。现实中的很多问题也适用于这个学习模型,例如图像分类、文本分类。图像分类中可以将一副图像看作一个包,图像可以分割成多个区块,这些区块可以作为多个示例,而一幅图像也可以对应很多的语义标签,例如海滩、云、大海等。文本分类中,每篇文章可以作为一个包,文章的每个段落可视为示例,而文章也可以被赋予多个主题。

专利文本一般具有固定格式,包括标题、摘要、权利要求、说明书、说明书附图等部分,这些部分正好可以作为多个示例,而这多个示例也组成了一个包,也就是一篇专利文本,该专利也对应一个或者多个标签,因此专利文本具有多示例多标签特征。如表 1 所示,可以将这篇专利看成一个包,专利名称、摘要等内容则是多个示例,经过 SooPAT 网站中 IPC 号检索,该专利的 IPC 分类号包含 D06(织物等的处理;洗涤;其他类不包括的柔性材料)、B32(层状产品)、C08(有机高分子化合物;其制备或化学加工;以其为基料的组合物)等,而通过专家对专利中各个示例的语义分析,可以发现该包中至少存在一个 D06、B32、C08 的 IPC 分类号的正例。

通过上述的描述,可以联想到将 MIML 机器学习模型运用到专利文本的 IPC 分类号的确定中来。要通过 MIML 模型进行专利分类,首先对已有 IPC 分类号的专利包进行训练,然后运用学习好的模型对未知分类号的专利数据进行 IPC 分类号预测,从而对未知 IPC 分类号的专利进行分类。

2.2 MIML 学习模型的数学描述

多示例多标签学习模型基于多标签学习以及多示例学习,它是一种较为一般的表现形式,包括了单示例单标签学习、多示例单标签学习、单示例多标签学习的各种情况,以上 3 种学习模型可以由多示例多标签学习退化得到,因此,多示例多标签学习具有普遍性、完整性等特点。

MIML 的数学形式可以表示为:令 X 表示示例空间,Y 表示标签空间,可以通过数据集$\{(X_1, Y_1), (X_2, Y_2), \cdots, (X_m, Y_m)\}$训练获得函数 $f: 2^X \to 2^Y$。其中,X_i 是一个包,用来描述一个真实对象,其由一组示例$\{x_{i1}, x_{i2}, \cdots, x_{in_i}\}$, $x_{ij} \in X(j = 1, 2, \cdots, n_i)$组成,而 Y_i 表示一组示例所对应的标签$\{y_{i1}, y_{i2}, \cdots, y_{il_i}\}$, $y_{ik} \in Y(k = 1, 2, \cdots, l_i)$。其中,$n_i$ 为描述第 i 个真实对象的示例个数,l_i 为第 i 个真实对象的标签个数。

2.3 MIML 学习模型的介绍

MIML 学习模型因其适用性并经过数十年的发展,已取得非常好的理论延伸,其理论上的创新主要集中在分类器的学习方法上。MIML 学习模型的类型主要有:第一类是基于正则化的思路[12-13],该思路需要确立优化模型和约束条件,并在此基础上进行求解;第二类是基于退化策略的思路[14],该思路主要以多示例或者多标签学习为桥梁,将 MIML 问题退化成传统监督

包翔，刘桂锋，崔靖华. 多示例多标签学习在中文专利自动分类中的应用研究[J]. 图书情报工作,2021,65(8):107 – 113.

表 1　MIML 结构:以专利文本为例

专利包中的示例	具体内容	通过专家判断所属 IPC 分类号
专利名称	一种环保聚氨酯革、制备方法及其制品	C08(有机高分子化合物;其制备或化学加工;以其为基料的组合物)
专利摘要	一种环保聚氨酯革,包括超纤无纺布、覆盖在所述无纺布上的底层,以及覆盖在所述底层上的面层、聚氯乙烯发泡层、水性粘接层,半聚氨酯革半成品发泡压花后、采用三版表处机进行表面处理,水性表处剂的上浆量为 40 – 60g/m^2,干燥温度和时间分别为 140 – 150℃,60 – 90s,表处后的革坯经揉纹、干燥、量尺后即得成品,所述面层中包括 4 层涂层,分别为:粘结层、第一中间层、第二中间层、表面层	D06(织物等的处理;洗涤;其他类不包括的柔性材料) B32(层状产品)
专利说明书	本发明涉及人工皮革领域,具体涉及一种环保聚氨酯革、制备方法及其制品。 本发明解决的技术问题在于,提供一种环保聚氨酯革,从源头消除有机溶剂污染,节约有机溶剂资源,消除安全隐患,改善工作环境,提升了半 PU 革的生态等级和国际市场竞争力。 采用水性材料替代溶剂型材料从源头消除有机溶剂造成的环境污染,节省大量有机溶剂资源,消除了火灾隐患,显著改善劳动者工作环境,有利于行业的可持续发展。 水性材料替代溶剂型材料,满足了欧盟生态半 PU 革要求,提高了半 PU 革产品生态等级和国际市场竞争力	D06(织物等的处理;洗涤;其他类不包括的柔性材料) C08(有机高分子化合物;其制备或化学加工;以其为基料的组合物)
专利权利要求书	一种环保聚氨酯革,包括超纤无纺布、覆盖在所述无纺布上的底层,以及覆盖在所述底层上的面层、聚氯乙烯发泡层、水性粘接层,其特征在于:半聚氨酯革半成品发泡压花后、采用三版表处机进行表面处理,水性表处剂的上浆量为 40 – 60g/m^2,干燥温度和时间分别为 140 – 150℃,60 – 90s,表处后的革坯经揉纹、干燥、量尺后即得成品,所述面层中包括 4 层涂层	D06(织物等的处理;洗涤;其他类不包括的柔性材料) B32(层状产品) C08(有机高分子化合物;其制备或化学加工;以其为基料的组合物)

学习问题;第三类则是借助其他方法解决 MIML 问题,例如神经网络[15 – 16]、梯度下降算法[17]等解决分类和优化问题。本文列举部分经典的 MIML 学习模型进行介绍与实验,如表 2 所示:

表 2　各种 MIML 学习模型的简单介绍

模型名称	原理	求解过程	优缺点
M3MIML[12]	基于正则化以及最大化间隔策略	为每个类别都假设一个线性模型,学习任务被表述为二次规划(QP)问题,并以对偶形式实现	直接利用了示例与标签之间的联系,不会遗漏信息,优化过程太多导致算法效率不高,特别是在训练集数量较多的情况下
MIML-BOOST[14]	基于退化策略	将多示例多标签转化为多示例单标签,利用 Boosting 方法对转化得到的多示例样本进行求解从而转化为传统的监督问题进行求解	算法简单,但是在转化过程中会遗漏相关信息
MIMLSVM[14]	基于退化策略	将多示例多标签转化为单示例多标签,利用 SVM 对转化得到的多标签问题进行分析从而转化为传统的监督问题进行求解	算法简单,时间效率高,但是在转化过程中会遗漏相关信息
MIMLRBF[15]	基于径向基(RBF)神经网络	输入层是包含示例的包,隐层是包聚类之后的聚类中心,通过是误差平方和最小化的方法来优化隐层与输出层的权值矩阵	直接建立示例和标签之间的联系,但是当训练数据有噪声或不易识别时,会导致网络性能的整体误差增大
MIMLNN[16]	基于反向传播(Back-propagation, BP)神经网络	包含两个阶段多层感知器(Multilayer Perceptron, MLP),并基于反向传播算法训练 MIML 模型	直接建立示例和标签之间的联系,并且考虑到标签与标签之间的相关性,但是算法中需要提前确定多个参数

MIML 机器学习模型已经被广泛用于文本分类中[11 – 18],在 MIML 研究中最常用的是 Reuters – 21578 文本数据[19],其主要作为标准测试数据集被用于 MIML 模型评价中,该数据文本分类数据集包含 2 000 个与 7 个标签关联的文档,每个包对应一个文档,通过滑动窗口技术将文档分割成多个示例,示例总数为 7 119个,采用词袋表示法提取 243 维特征向量表示示例。Y. Yang 等[20]建立 WKG Game-Hub,将 MIML 用于网络游戏的角色分类中,其文本语料从“王者荣耀”游戏中心收集,由 13 750 篇文章组成,共有 1 744 个概念标签。总体来说,目前还尚未有文献将 MIML 学习模型用于专利文本的分类中。

2.4　模型效果评估指标

MIML 模型的学习效果评估一般通过 Average precision、Hamming Loss、Ranking Loss、One Error、Coverage、Training time (训练时间)等 6 个指标对两个未知的参数进行确定以及之后的性能评价[12],其中,Average precision 为分类的准确率,Training time 为训练 MIML 模型所耗费的时间。

Hamming Loss 指标反映的是样本在某一个标签上的误分类程度,包含两个情况,一个是相关标签没有出现在预测的标签集合,另一个是无关标签出现在预测的标签集合中,因此,该指标取值越小则学习模型越优。其计算公式如下:

$$Hamming\ Loss = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \Delta Z_i|}{M} \quad 公式（1）$$

其中，m 代表样本个数，M 代表所有标签的总个数，Y_i 代表样本 i 实际标签的个数，Z_i 代表样本 i 预测标签的个数，Δ 表两个集合的异或操作。

Ranking Loss 评价指标用于评价在样本的标签排序序列中出现排序错误的数值，在排序序列中无关标签排序优先于相关标签，同样的，该指标取值越小则学习模型越优。其计算公式如下：

$$Ranking\ Loss = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{|Y_i||\bar{Y}_i|}|\{(y_1,y_2)|f(x_i,y_1)\leq f(x_i,y_2),(y_1,y_2)\in Y_i * \bar{Y}_i\}| \quad 公式（2）$$

其中，$\bar{Y}_i$ 是 Y_i 相对于所有类别标签集合的补集。$f(x,y)$ 为实值函数，返回 x 的标签 y 的置信度。

One Error 评价指标是用于评价样本标签排序集合中，排在最前面的标签不属于相关标签集合的指标，该指标取值越小则学习模型越优。其计算公式如下：

$$One\ error = \frac{1}{m}\sum_{i=1}^{m}\{[arg_{y\in Y}\ maxf(x_i,y)]\notin Y_i\} \quad 公式（3）$$

Coverage 评价指标用于评价在样本的标签排序集合中，遍历所有相关标签需要的搜索深度，该指标取值越小则学习模型越优。其计算公式如下：

$$Coverage = \frac{1}{m}\sum_{i=1}^{m} max_{y\in Y}\ rank_{f(x_i,y)} - 1 \quad 公式（4）$$

$rank_{f(x_i,y)}$ 返回由 $f(x,.)$ 按降序排列的标签 y 的秩。

3 实验结果与分析

3.1 实验数据与实施流程

本文从上海知识产权公共服务平台的中国专利数据库中选取水处理技术领域专利文献作为语料库，这些专利数据主要包含分类号、标题、摘要、主权项等内容。根据 Soopat 网站中的 SooPAT IPC 检索结果，专利的主要分类号以及对应主题分别是：B01D（分离）；C02F（水、废水、污水或污泥的处理）；D06（织物等的处理、洗涤、其他类不包括的柔性材料），每类专利文献 250 篇，其中有 60 篇左右的专利包含有 B01D、C02F、D06 中两个及以上的分类号。

本实验的方法流程如图 2 所示，包含数据库构建、文本预处理与向量化、模型训练与参数调节、模型分类效果评估等方面。具体的数据库构建包括对含有标签的专利数据按照一定的比例进行训练集和测试集选取；文本预处理和向量化则包括分词、停用词、词性标注后删除某些词性的词语，建立基于 TF-IDF 的向量空间模型；模型训练与参数调节则是选取本文第二节介绍的经典 MIML 模型，配合参数调节的方法，达到该模型的最优效果；模型评估则是对模型进行效果评估，主要指标包含 Average precision、Ranking Loss、Hamming Loss、One Error、Coverage 等上文介绍的指标。

将每一篇专利看作一个包，专利的标题和摘要当作包中的两个示例。具体的，共选取 200 篇专利文本作为实验数据，其中具有多个标签的包占比约为 30%，平均每个包有 1.29 个标签。在分词阶段，本实验采用 jieba 中文分词的.NET 版本并通过精确分词模式来实现，在特征选择阶段，选取了前 1 000 个 TF * IDF 值对应的特征词作为数据的索引词[21]。本文实验所用处理器参数为：Intel(R) Core(TM) i5 – 7500 CPU @3.40GHz，内存 4GB，64 位操作系统，基于 x64 处理器，实验所用的软件是 Matlab R2018a。

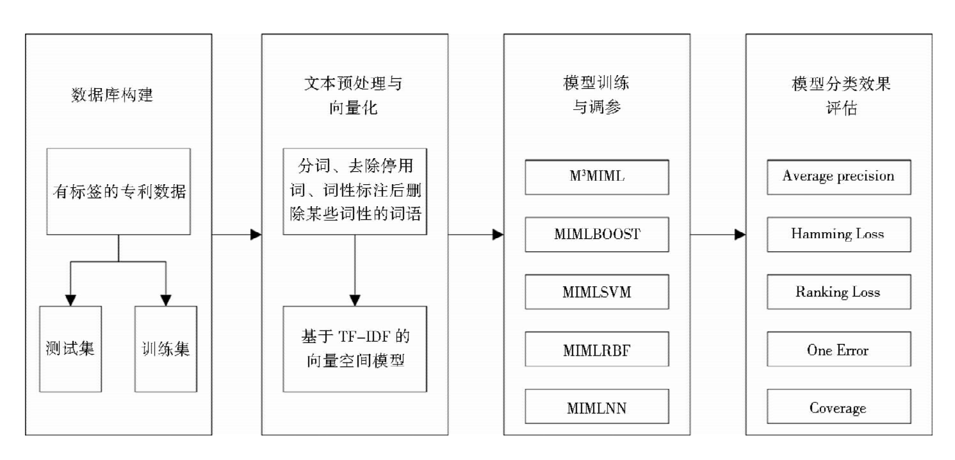


图 2 方法实施流程

3.2　分类结果与分析

为了验证各个模型的分类效果以及模型的适用性,本文拟采用 10 次 N 折交叉验证的方法,N 的取值为 2 - 10,并且计算 10 次测试的平均值作为指标参数。为了获得更好的实验效果,综合各种文献数据以及实验数据,将 MIML 各种的模型中的实验参数调整如下:M3MIML 中线性模型选择 SVM 线性判断模型,损失函数阈值为 0.01,MIMLBOOST 同样选择 SVM 线性判断模型,循环次数 rounds 设置为 100,MIMLSVM 选择 SVM 中的 RBF 判断模型,比例参数 ratio 设置为 0.2,MIMLRBF 则将比例参数 ratio 设置为 0.1,MIMLNN 的网络判断阈值设置为 0.5。综合考虑到 N 的不同取值下指标的变化情况,选取 N = 3、4、5 的评价结果显示并讨论,各个 MIML 模型在水处理中文专利的分类结果分析如表 3 - 表 5 所示,评价指标包括 Average precision、Ranking Loss、Hamming Loss、One Error、Coverage、Training time。

表 3　各个 MIML 模型通过三折交叉验证的方法的结果

模型名称	Aver precision	Ranking Loss	Hamming Loss	One Error	Coverage	Training time/s
M3MIML	0.817 5 ±0.004	0.232 8 ±0.009	0.264 4 ±0	**0.327 6 ±0.017**	**0.724 1 ±0**	**10.21 ±1.110**
MIMLBOOST	**0.800 6 ±0.003**	**0.232 2 ±0.036**	**0.291 7 ±0.006**	**0.392 9 ±0**	**0.696 5 ±0.054**	**3.364 ±1.186**
MIMLSVM	**0.809 9 ±0.003**	**0.210 6 ±0.004**	0.269 1 ±0.005	0.386 1 ±0.006	**0.649 1 ±0.006**	**0.188 6 ±0.008**
MIMLRBF	**0.828 8 ±0.010**	0.219 2 ±0.005	**0.262 3 ±0.048**	0.333 7 ±0.023	0.823 3 ±0.073	0.382 8 ±0.055
MIMLNN	0.791 7 ±0.024	0.267 9 ±0.071	0.333 3 ±0.071	0.410 7 ±0.054	0.750 0 ±0.107	0.272 1 ±0.008

表 4　各个 MIML 模型通过四折交叉验证的方法的结果

模型名称	Aver precision	Ranking Loss	Hamming Loss	One Error	Coverage	Training time/s
M3MIML	0.842 8 ±0.010	0.263 7 ±0.006	0.232 2 ±0.018	0.256 5 ±0.029	0.852 3 ±0.005	19.25 ±0.781
MIMLBOOST	0.781 4 ±0.016	0.356 1 ±0.023	0.256 0 ±0.031	0.392 9 ±0	**0.417 8 ±0.037**	2.551 ±0.081
MIMLSVM	0.831 1 ±0.014	0.209 9 ±0.028	0.310 3 ±0.007	0.349 6 ±0.031	0.559 0 ±0.006	0.378 1 ±0.044
MIMLRBF	**0.888 9 ±0.036**	**0.154 8 ±0.048**	**0.230 1 ±0.040**	**0.214 3 ±0.071**	0.642 9 ±0.024	0.484 4 ±0.016
MIMLNN	0.756 0 ±0.060	0.313 9 ±0.044	0.325 4 ±0.008	0.488 1 ±0.012	0.885 3 ±0.067	**0.280 4 ±0.002**

表 5　各个 MIML 模型通过五折交叉验证的方法的结果

模型名称	Aver precision	Ranking Loss	Hamming Loss	One Error	Coverage	Training time/s
M3MIML	0.808 8 ±0.020	0.264 7 ±0	**0.284 3 ±0.010**	0.353 0 ±0.059	0.853 5 ±0	16.39 ±3.984
MIMLBOOST	0.769 6 ±0.010	0.338 3 ±0.044	0.382 4 ±0.069	0.441 4 ±0.030	1 ±0.059	3.684 ±0.108
MIMLSVM	0.821 1 ±0.002	**0.205 9 ±0.030**	0.294 1 ±0.020	0.352 9 ±0	**0.617 7 ±0.009**	0.445 3 ±0.008
MIMLRBF	**0.840 7 ±0.017**	0.220 6 ±0.044	0.303 9 ±0.001	**0.323 5 ±0.030**	0.676 5 ±0.206	0.453 2 ±0.031
MIMLNN	0.790 9 ±0.022	0.235 3 ±0.029	0.284 3 ±0.049	0.382 4 ±0.029	0.794 1 ±0.029	**0.272 7 ±0.001**

表 3 - 表 5 中粗体的数据为 N 折交叉验证下的最优指标,通过分析可以得到以下结论:

(1)MIML 模型大多能准确地对专利进行分类,所有模型的分类精确度都在 80% 左右,这说明 MIML 学习模型具有较高的准确性,因此适用于确定中文专利 IPC 号的工作。

(2)随着 N 的增加,各个模型样本的训练时间总体都在增加,这是由于训练样本的增加所致,但是其余的性能指标并未随着 N 的增加而变优。总体来说,各个模型都在四折交叉验证的时候取得较好的指标参数,这可能是因为若 N 小于 4 的时候,训练样本的数量不多,未能充分进行模型的训练;而当 N 大于 4 时,训练样本过多,容易引起模型的过拟合现象,从而导致模型的泛化能力变差。

(3)从模型选择上考虑时,发现选取不同的 N 折交叉验证方法时,MIMLRBF 模型的 Average precision

总是最高的，而且其他指标虽然并没有都是最优，但是与最优的指标相差并不多，且训练效率仅次于 MIMLNN 模型，明显好于 M3MIML、MIMLBOOST 模型。MIMLRBF 模型由于使用神经网络结构进行问题求解，优势在于在输入层与隐含层之间的聚类过程和隐含层与输出层的优化过程中，示例和标签之间的连接都是明确的，实践也说明了 MIMLRBF 能准确、快速地解决问题。

综上所述，MIML 模型较之于传统的监督学习模型有着较大的优势，因为相较于传统分类模型只考虑单示例单标签的思想，MIML 模型既充分考虑到了专利文本的多示例的结构属性，又考虑了专利的多标签属性，可以多角度地选取数据作为专利文本分类的依据，分类的结果自然也会更加精确。类似地，本文也可以将权利要求书、说明书等专利信息作为示例进行训练，而且本文也考虑不同的 N 折交叉验证方法对实验结果的影响，从而可以获得更加科学的训练数据和测试数据的配比。实践证明了四折交叉验证时的分类效果最好，即将训练集比例设置为 75%，将测试集比例设置为 25%；并且推荐使用 MIMLRBF 模型进行中文专利的分类，也提示要选择示例和标签之间具有明确连接的模型进行中文专利的分类工作。

4 结语

本文充分考虑到专利文本的结构特点以及其固有的多标签属性，将 MIML 机器学习模型运用在中文专利的分类中，实验指标表明 MIML 模型能较为准确、快速地实现中文专利的分类，使得大规模进行自动的专利 IPC 分类号的确定成为可能，大大减少了人工标注的效率低下、受标注人知识结构影响等弊端。只需少量的样本数据，就能实现大规模的专利分类，是人工智能技术在图书情报领域内的积极尝试。在此也对 MIML 模型运用的优势与不足进行总结与展望：①MIML模型能适用于实际专利分类现状，特别是只有少量标签数据的情形下，本文提出的思路可以拓展专利分类的应用范围，辅助确定大量未标注专利的多个标签的类别。②通过 MIML 模型进行中文专利的分类实验可以得知，很多 MIML 模型的训练效率都非常高，这也为高效、准确地专利分类提供了思路。③实验中 MIML 模型已经提前通过实验的方法确定了实验参数，因此专利分类的结果较为准确。但是如果实际情况中，在不知道训练集、测试集的具体情况时，需要提前确定 MIML 模型的部分参数，因此，其分类的准确性、效率将会受到影响。

针对上文的实验与分析，今后还要对以下问题进行研究：①本实验选取的专利测试样本有限，并且这些专利对应的标签个数很少，若是在实际情况中，需要进行标注的专利数量庞大，有时还对应更多的标签，如何选择专利文本的特征、分词方法和 MIML 模型是研究的关键方向。②本实验将只选取标题和摘要作为示例进行训练，但是专利文本中说明书和权利要求书也存在这大量技术信息，如何将这些内容放进 MIML 模型中，并寻找哪些示例的组合拥有较高的分类准确率也是需要思考的问题。③传统 MIML 模型的参数繁多且难以确定，并且参数可能对算法分类准确率的影响相当大，如何确定一个快速、精确的参数估计方法，并将它广泛应用于专利分类的研究也是下一步需要解决的问题。

参考文献：

[1] 高莉. 科技创新市场化的专利制度回应[J]. 江苏大学学报(社会科学版)，2017，19(1)：63–69.

[2] 吕璐成，韩涛，周健，等. 基于深度学习的中文专利自动分类方法研究[J]. 图书情报工作，2020，64(10)：75–85.

[3] 胡杰，李少波，于丽娅，等. 基于卷积神经网络与随机森林算法的专利文本分类模型[J]. 科学技术与工程，2018，18(6)：268–272.

[4] 张群，王红军，王伦文. 词向量与 LDA 相融合的短文本分类方法[J]. 现代图书情报技术，2016，32(12)：27–35.

[5] 温超东，曾诚，任俊伟，等. 结合 ALBERT 和双向门控循环单元的专利文本分类[J]. 计算机应用，2021，41(2)：407–412.

[6] 余本功，张培行. 基于双通道特征融合的 WPOS-GRU 专利分类方法[J]. 计算机应用研究，2020，37(3)：655–658.

[7] GOMEZ J. Analysis of the effect of data properties in automated patent classification[J]. Scientometrics，2019，121(3)：1239–1268.

[8] 胡学钢，杨恒宇，林耀进，等. 基于协同过滤的专利 TRIZ 分类方法[J]. 情报学报，2018，37(5)：512–518.

[9] LI S，HU J，CUI Y，et al. DeepPatent：patent classification with convolutional neural networks and word embedding[J]. Scientometrics，2018，117(2)：721–744.

[10] 周成，魏红芹. 专利价值评估与分类研究——基于自组织映射支持向量机[J]. 数据分析与知识发现，2019，3(5)：117–124.

[11] LU Y，XIONG X，ZHANG W，et al. Research on classification and similarity of patent citation based on deep learning[J]. Scientometrics，2020，123(2)：813–839.

[12] ZHANG M L, ZHOU Z H. M3MIML: A maximum margin method for multi-instance multi-label learning[C]//Eighth IEEE international conference on data mining. Los Alamitos: IEEE Computer Society, 2008:688－697.

[13] ZHOU Z H. A framework for machine learning with ambiguous objects[C]// 5th international conference on active media technology. Berlin: Springer-Verlag, 2009:6.

[14] ZHOU Z H, ZHANG M L. Multi-instance multi-label learning with application to scene classification[C]// Advances in neural information processing systems. Cambridge: Neural information processing systems foundation, 2006: 1609－1616.

[15] ZHANG M L, WANG Z J. MIMLRBF: RBF neural networks for multi-instance multi-label learning[J]. Neurocomputing, 2009, 72(16－18):3951－3956.

[16] CHEN Z, CHI Z, FU H, et al. Multi-instance multi-label image classification: a neural approach[J]. Neurocomputing, 2013, 99(1):298－306.

[17] HUANG S J, GAO W, ZHOU Z H. Fast multi-instance multi-label learning[J]. IEEE transactions on pattern analysis and machine intelligence, 2014, 41(11):1868－1874.

[18] 严考碧,李志欣,张灿龙. 基于主题模型的多示例多标记学习方法[J]. 计算机应用,2015,35(8):2233－2237.

[19] SEBASTIANI F. Machine learning in automated text categorization [M]. New York: ACM, 2002.

[20] YANG Y, WU Y F, ZHAN D C, et al. Complex object classification: a multi-modal multi-instance multi-label deep network with optimal transport[C]//The 24th ACM SIGKDD international conference. New York: Assoc Computing Machinery, 2018:2594－2603.

[21] 包翔,刘桂锋,杨国立. 基于多示例学习框架的专利文本分类方法研究[J]. 情报理论与实践,2018,41(11):144－148.

作者贡献说明：
包翔：整体构思，论文写作，实验操作；
刘桂锋：整体构思，论文修改；
崔靖华：数据分析。

Application of Multi Instance Multi Label Learning in Chinese Patent Automatic Classification

Bao Xiang[1] Liu Guifeng[1] Cui Jinghua[2]
[1] Institute of Science and Technology Information, Jiangsu University, Zhenjiang 212013
[2] School of Information Management, Nanjing University, Nanjing 210093

Abstract: **[Purpose/significance]** In order to achieve rapid classification in a large number of Chinese patents to meet the requirements of patent examination and intelligence analysis. **[Method/process]** Combined with the inherent format of patent text and the fact that there are multiple classification numbers, this paper applied multi-instance multi-label learning to automatic patent classification. Firstly, several classical multi-instance multi-label learning methods were introduced, and then these methods were applied to determine IPC number of Chinese patent. **[Result/conclusion]** It is experimentally demonstrated that the multi-instance multi-label learning methods are suitable for patent automatic classification, according to average precision, hamming loss, ranking loss, one error, coverage, training time, it is found that MIMLRBF can be used to determine the IPC number of Chinese patents quickly and accurately, which provides a new perspective for classifying large-scale patents.

Keywords: patent classification IPC multi-instance multi-label